

Panel data segmentation under finite time horizon

Leonid Torgovitski^{*†}

*Mathematical Institute, University of Cologne,
Weyertal 86-90, 50931, Cologne, Germany*

Abstract

We study the nonparametric change point estimation for common changes in the means of panel data. The consistency of estimates is investigated when the number of panels tends to infinity but the sample size remains finite. Our focus is on weighted denoising estimates, involving the group fused LASSO, and on the weighted CUSUM estimates. Due to the fixed sample size, the common weighting schemes do not guarantee consistency under (serial) dependence and most typical weightings do not even provide consistency in the i.i.d. setting when the noise is too dominant.

Hence, on the one hand, we propose a consistent covariance-based extension of existing weighting schemes and discuss straightforward estimates of those weighting schemes. The performance will be demonstrated empirically in a simulation study. On the other hand, we derive sharp bounds on the change to noise ratio that ensure consistency in the i.i.d. setting for classical weightings.

Keywords: Panel data, Change point estimation, Segmentation, Nonparametric, CUSUM, Total variation denoising, LASSO, Serial dependence

^{*}E-mail: ltorgovi@math.uni-koeln.de

[†]Research partially supported by the Friedrich Ebert Foundation, Germany.

¹© 2015, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International <http://creativecommons.org/licenses/by-nc-nd/4.0/>

²The final article is published under <http://dx.doi.org/10.1016/j.jspi.2015.05.007>

1 Introduction

The aim of this paper is to study the estimation of changes in the context of *panel data*. We focus on *common changes*, i.e. changes that occur simultaneously in many panels (but not necessarily in all) at the same time points and we consider an asymptotic framework where the number d of panels tends to infinity but the panel sample size n is fixed.

The analysis of change point estimation in panel data is subject of intensive research (in particular in econometrics) and, as discussed in Bai (2010), dates back at least to the works of Joseph and Wolfson (1992, 1993). However, the setting $d \rightarrow \infty$, which we are looking at, is generally not studied much in the literature concerning change point analysis and the settings $n \rightarrow \infty$ or $n, d \rightarrow \infty$ are far more established.

For the classical setting of $n \rightarrow \infty$ we refer to Csörgő and Horváth (1997). In the context of panel data especially the setting $n, d \rightarrow \infty$ is quite popular (cf., e.g., Bai (2010), Horváth and Hušková (2012) and Kim (2014)). Nevertheless, the assumption $d \rightarrow \infty$ and n fixed is also quite natural (cf., e.g., Bai (2010), Bleakley and Vert (2011a), Hadri et al (2012) and also Peštová and Pešta (2015)). It reflects the situation where the amount of panels, i.e. the dimensionality, is much larger than the sample size.

Bai (2010) and Bleakley and Vert (2011a) mention important applications in finance, biology and medicine where in particular the framework of common changes is appropriate: In finance such changes may occur simultaneously across many stocks e.g. due to a credit crisis or due to tax policy changes. In biology and medicine relevant applications are in the study of genomic profiles within classes of patients. As mentioned in Bleakley and Vert (2011a) the latter example fits particularly well in the n fixed and $d \rightarrow \infty$ framework because the length of panels in genomic studies is fixed but the amount of panels can be increased by raising the number of patients.

The body of literature related to change point estimation (and detection) is huge. Hence, we do not attempt to summarize it here and refer the reader instead to the reviews in Jandhyala et al (2013), Aue and Horváth (2013), Frick et al (2014) and Horváth and Rice (2014). Change point analysis in the $d \rightarrow \infty$ and n fixed setting goes at least back to the (aforementioned) papers by Bleakley and Vert (2010, 2011a) and by Bai (2010). Therein estimation of common changes is studied independently from different perspectives. However, as we will see, the setups of Bleakley and Vert (2010, 2011a) and of Bai (2010) are closely related¹.

Bai (2010) considered a *least squares* estimate for independent panels of linear time series under a single change point assumption and Bleakley and Vert (2011a) developed a weighted *total variation denoising* approach for the multiple change point scenario. Furthermore, Bleakley and Vert (2011a) proposed a computationally efficient algorithm and implemented it in a convenient MATLAB package GFLseg² which we also used in some of our simulations.

In this article we study consistency properties, in particular what we define as *perfect estimation*³, for the denoising estimate and for the weighted CUSUM (cumulative sums) estimate under weak dependence. Both types of estimates depend on certain weighting

¹Notice that Bleakley and Vert (2011a) is a revised version of Bleakley and Vert (2010). Hence, we will mostly refer to the more recent article.

²Download is available at <http://cbio.ensmp.fr/GFLseg> and is licensed under the GNU General Public License.

³See Subsection 2.2.3 and (2.20) below.

schemes w . Two schemes, w^{simple} and w^{standard} , were already considered by Bleakley and Vert (2011a) for the denoising approach in the n fixed and $d \rightarrow \infty$ setting (cf. Subsection 2.2.2 for the precise definition). They showed that w^{standard} ensures perfect estimation and therefore has better consistency properties for $d \rightarrow \infty$ than w^{simple} does. (Notice that Bai (2010) showed perfect estimation for the least squares estimate, which corresponds to the weighted CUSUM estimate with w^{standard} .)

We pick up the ideas of Bleakley and Vert (2011a) and extend them in various directions which will shed some new light on weighting schemes in general. First, we will emphasize the connection between the total variation denoising approach and the weighted CUSUM estimates. Notice that Bleakley and Vert (2011a) assumed independent panels of independent Gaussian observations. We continue by showing that their consistency results hold true under much weaker distributional assumptions, e.g. for panels of non-Gaussian time series with common factors. This is important since many datasets are neither Gaussian nor independent. An implication of our results is that w^{standard} generally does not provide consistency for panels of time series and therefore does not ensure perfect estimation under dependence.

As a solution, we propose a modified weighting scheme w^{exact} , which is a generalization of w^{standard} , that takes the covariance structure within panels into account. We show that this is the only choice that may generally ensure perfect estimation and derive quite mild conditions under which w^{exact} indeed ensures this property. In a detailed simulation study we confirm our results and demonstrate the gain in accuracy of w^{exact} . Moreover, we show that our approach outperforms the classical schemes even in random change point settings and for rather moderate dimensions. In practice, the weights w^{exact} have to be estimated. Therefore, we discuss feasible approaches and show their applicability in simulations.

Complementary to the study of perfect estimation, we investigate consistent estimation for a further class of weights w^{weighted} , which contains w^{simple} and w^{standard} as special cases, and characterize changes which are (not) correctly estimated as $d \rightarrow \infty$.

1.1 Basic setup

We observe d panels $\{Y_{i,k}\}_{i=1,\dots,n}$ for $k = 1, \dots, d$ in a *signal plus noise* model where

$$Y_{i,k} = m_{i,k} + (\varepsilon_{i,k} + \gamma_k \zeta_i). \quad (1.1)$$

Here, $\{m_{i,k}\}_{i,k \in \mathbb{N}}$ is an array of deterministic signals and $\{\varepsilon_{i,k}\}_{i,k \in \mathbb{N}}$ is an array of random centered noises. The $\{\zeta_i\}_{i \in \mathbb{N}}$ are the so-called *common factors* which are assumed to be random, centered and independent of $\{\varepsilon_{i,k}\}_{i,k \in \mathbb{N}}$. Their effect on the k -th panel is quantified via the deterministic *factor loadings* $\gamma_k \in \mathbb{R}$.

We assume a (multiple) common change points scenario given by

$$m_{i,k} = \begin{cases} \mu_{1,k}, & i = 1, \dots, u_1, \\ \mu_{2,k}, & i = u_1 + 1, \dots, u_2, \\ \dots, & \dots, \\ \mu_{P+1,k}, & i = u_P + 1, \dots, n, \end{cases} \quad (1.2)$$

where we call $u_1, \dots, u_P \in \mathbb{N}$ change points. The $\mu_{j,k} \in \mathbb{R}$, $j = 1, \dots, P+1$, describe the piecewise constant signals in each panel, i.e. the means of the observations. In other

words the means jump simultaneously from levels $m_{u,k}$ to levels $m_{u+1,k}$ in all panels $k = 1, \dots, d$ at change points $u \in \{u_1, \dots, u_P\}$. However, we do not require $m_{u,k} \neq m_{u+1,k}$ to hold for all $k = 1, \dots, d$, i.e. the changes do not have to occur in all panels. Later on we will impose more specific conditions on the average magnitude of changes.

Subsequently, we assume that $n \geq 3$ since otherwise the model (1.2) is not reasonable because for $n = 1$ the model may not contain any change and for $n = 2$ it holds trivially that $P = 1$ with $u_1 = 1$.

1.2 Notation

We follow the compact matrix notation of Bleakley and Vert (2011a) and represent the model (1.2) as

$$Y = \mathcal{M} + E,$$

with a deterministic matrix \mathcal{M} of means with $\mathcal{M}_{i,k} = m_{i,k}$ and a random matrix of errors E with $E_{i,k} = \varepsilon_{i,k} + \gamma_k \zeta_i$. Now, let X be any $n \times d$ matrix. To shorten the notation we write $X_{\bullet,j}$ for $[X_{1,j}, \dots, X_{n,j}]^T$ and $X_{i,\bullet}$ for $[X_{i,1}, \dots, X_{i,d}]$. For example $Y_{\bullet,k}$ represents the k -th panel and a common change at u corresponds to

$$\Delta = \mathcal{M}_{u+1,\bullet} - \mathcal{M}_{u,\bullet} \neq 0. \quad (1.3)$$

$\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_2$ stands for the Euclidean norm. We simply write $\|\cdot\|$ for the former when no confusion is possible, unless it is stated otherwise.

We will consider functions $f(i)$ with a discrete support $i = 1, \dots, n-1$ and say that a function f is convex (or concave) if this holds true for the linear interpolation of points $f(i)$ on the interval $[1, n-1]$. Subsequently, we mean by $\arg \max$ the whole set of points at which the maximum is attained.

The paper is organized as follows. In Section 2 we discuss segmentation of panel data. In Subsection 2.1 we introduce the concept of the denoising segmentation approach in general and then we turn to the single change point scenario in Subsection 2.2. First, we clarify the selection of a certain regularization parameter and continue to discuss the relation to a class of weighted CUSUM estimates in Subsection 2.2.1. Common weighting schemes are presented in Subsection 2.2.2. In Subsection 2.2.3 we analyze the segmentation procedures with respect to different weighting schemes and propose a generalization of existing approaches. Subsequently, we discuss estimates of the generalized weighting scheme in Subsection 2.2.4. In Section 3 we confirm our theoretical results in a simulation study. Finally, we provide a short summary of the paper in Section 4 and all proofs are postponed to Section 5.

2 Segmentation of panel data

We start with a description of the denoising approach to change point estimation of Bleakley and Vert (2011a). For an overview of the related literature we refer to the references therein.

2.1 Total variation denoising estimates

The total variation denoising approach to segmentation is to solve the convex minimization problem⁴

$$\underset{U \in \mathbb{R}^{n \times d}}{\text{Minimize}} \quad \frac{1}{2} \|Y - U\|_F^2 + \lambda \times \text{totvar}(U) \quad (2.1)$$

for an appropriate regularization parameter $\lambda \geq 0$ under a weighted total variation penalty term

$$\text{totvar}(U) = \sum_{i=1}^{n-1} \frac{\|U_{i+1,\bullet} - U_{i,\bullet}\|_2}{w(i, n)} \quad (2.2)$$

with positive, *position dependent* weights $w(i, n) > 0$. We denote the solution of (2.1) by $\hat{U}(\lambda)$ and each column $\hat{U}_{\bullet,k}$ represents the best piecewise constant fit to the panel $Y_{\bullet,k}$ with respect to (2.1). Each change in those fits, in the sense of $\hat{U}_{u+1,\bullet} \neq \hat{U}_{u,\bullet}$, is therefore assumed to identify a common change across panels at time point u . Hence, the set \mathcal{E} of estimated change points is given by

$$\mathcal{E}(\lambda) = \left\{ u \mid \hat{U}_{u,\bullet}(\lambda) \neq \hat{U}_{u+1,\bullet}(\lambda) \right\}. \quad (2.3)$$

The penalty term $\text{totvar}(U)$ is designed in such a way that $\hat{U}(\lambda)$ has for $\lambda > 0$ a tendency to reduce the cardinality of (2.3), i.e. to reduce the amount of identified change points. Hence, \mathcal{E} has a tendency to become smaller as λ increases.

Two extreme cases give some insight: For $\lambda \uparrow \infty$ the penalty term $\text{totvar}(U)$ dominates the minimization and forces the minimizer \hat{U} to be constant across rows, i.e. we obtain $\mathcal{E} = \emptyset$ and no change points are identified by this procedure at all. In contrast to this, if $\lambda = 0$ then $\hat{U}(0) = Y$ and therefore $\mathcal{E}(0) = \{u \mid Y_{u,\bullet} \neq Y_{u+1,\bullet}\}$, i.e. the number of estimated changes corresponds to the number of different consecutive rows of Y . Hence, if e.g. all rows are unique then each point $i = 1, \dots, n-1$ is identified as a change point.

2.2 Single change point scenario

Following Bleakley and Vert (2011a) we will at first restrict our considerations to the single change point scenario⁵.

Assumption 2.1. We consider a single change point scenario with a change at some time point $u \in \{1, \dots, n-1\}$ where $n \geq 3$.

We need to clarify the selection of λ for (2.1). In the single change point setup we aim to select λ as large as possible such that the set \mathcal{E} of change points contains only one change point⁶ in which case $\mathcal{E} = \{\hat{u}\}$ and \hat{u} denotes the *denoising estimate* for u .

⁴The objective function in (2.1) is strictly convex, as a sum of convex functions and due to the strict convexity of the mapping $U \mapsto \|Y - U\|_F^2$. Moreover, we may restrict the minimization to a compact subset. Therefore, a unique solution exists for any $\lambda \geq 0$.

⁵However, as will be shown in the simulations, our findings do have practical implications on the multiple change point scenario as well which is why we stated the general model in (1.2).

⁶For the single change point scenario Bleakley and Vert (2011a, in their software GFLseg) perform a dichotomic search to find the “first” λ such that \mathcal{E} contains only one element.

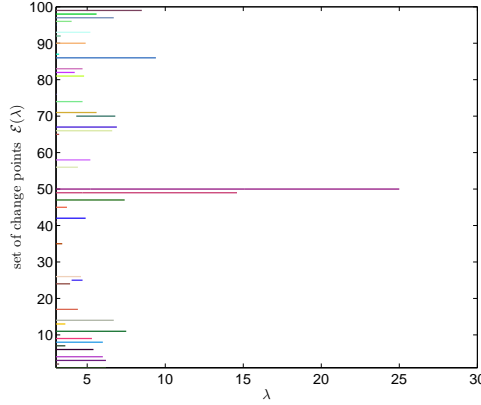


Figure 1: The set $\mathcal{E}(\lambda)$, as defined in (2.3), along the λ regularization path for $n = 100$, $d = 50$, independent panels with independent standard Gaussian noise $\{\varepsilon_{i,k}\}$ and without common factors, i.e. with $\gamma_k = 0$; the true change point lies at $u = 50$ with means $\mu_{1,k} = 0$ and $\mu_{2,k} = 1$, i.e. $\Delta_k = 1$, for all panels $k = 1, \dots, d$. Notice that a change point at $i = 25$ is only estimated for λ between 4 and 4.7. (The computations are performed using the MATLAB package GFLseg.)

Heuristically, this forces \hat{u} to be the most reasonable selection of exactly one common change point according to the penalty term $\text{totvar}(U)$.

As will be discussed in Proposition 5.1, one can identify under mild assumptions a random interval such that any $\lambda \in (\lambda_{\min}, \lambda_{\max})$, with $0 \leq \lambda_{\min} < \lambda_{\max}$, yields the same estimate \hat{u} and such that any $\lambda \in [\lambda_{\max}, \infty)$ yields $\mathcal{E} = \emptyset$, i.e. no estimate. Thus, in the following, we tacitly assume that we select any $\lambda \in (\lambda_{\min}, \lambda_{\max})$ in which case the corresponding estimate \hat{u} is unambiguous.

Notice that generally the number of estimated change points does not necessarily decrease monotonously in λ for $d > 1$ (cf. Figure 1 and Section 4 of Bleakley and Vert (2011a)) and, additionally, it seems not clear whether any parameter λ that identifies only one change point yields the same estimate.

Remark 2.2. A parameter λ that yields only one change point does not always exist. (E.g., it holds $\mathcal{E} = \emptyset$ for any λ if all entries of Y are equal.) We will consider situations where such cases do not occur with probability tending to 1 as $d \rightarrow \infty$.

Remark 2.3. For selection of some reasonable λ in the case of multiple changes, in particular if the number of changes is unknown in advance (which is a more realistic scenario), we refer to Bleakley and Vert (2011a) and to the references therein.

2.2.1 Relation to weighted CUSUM

In the next proposition we observe, based on Proposition 5.1, the relation of the estimate \hat{u} from the denoising approach to a well-known weighted CUSUM estimate⁷

$$\hat{u}_\star \in \arg \max_{i=1, \dots, n-1} t(i), \quad t(i) = w^2(i, n) \left(\sum_{k=1}^d \left| \sum_{j=1}^i (Y_{j,k} - \bar{Y}_{n,k}) \right|^2 \right). \quad (2.4)$$

⁷ \hat{u}_\star is usually defined as the smallest element in $\arg \max$. Here, we allow \hat{u}_\star to be any element in $\arg \max$.

Proposition 2.4. *Under Assumption 2.1 and given that $t(i)$ has a unique maximum, it holds that $\hat{u} \equiv \hat{u}_\star$ if we use the same weighting w for the denoising and the CUSUM estimates.*

Note that this connection holds true only in case of a single change point whereas otherwise denoising and CUSUM estimates differ. Therefore, recall that the denoising segmentation approach yields p distinct change point estimates in case of p change points.

Proposition 2.4 allows us to study denoising and CUSUM estimates simultaneously if we tacitly exclude the case of non unique maxima of $t(i)$ from our considerations⁸. This is not a problem, since we will focus mostly on situations where this case does not occur with probability tending to 1 as $d \rightarrow \infty$.

2.2.2 Common weighting schemes

Bleakley and Vert (2011a) already studied the weightings

$$w^{\text{simple}}(i, n) = 1, \quad w^{\text{standard}}(i, n) = ((i/n)(1 - i/n))^{-1/2} \quad (2.5)$$

for the denoising estimate \hat{u} with respect to $d \rightarrow \infty$ and the latter has also been studied by Bai (2010) for the least-squares estimate. Both schemes can be considered as natural and are reasonable for the denoising and for the CUSUM estimate as well. The former, w^{simple} , appears to be the first choice from the point of view of the denoising approach. In fact, Bleakley and Vert (2010) started with this case and studied w^{standard} later in Bleakley and Vert (2011a). On the other hand, the latter weighting, w^{standard} , appears to be the natural choice from the CUSUM point of view, because it can be derived via a maximum-likelihood or a least-squares approach. Both weights are special cases of the following parametrized scheme

$$w^{\text{weighted}}(i, n) = ((i/n)(1 - i/n))^{-\gamma}, \quad 0 \leq \gamma \leq 1/2. \quad (2.6)$$

These weights are quite popular in the field of change point analysis. Asymptotic properties are well studied for $n \rightarrow \infty$ for testing with weighted CUSUM statistics⁹, via $\max_{i=1, \dots, n-1} t^{1/2}(i)$, or for estimating changes via $\arg \max_{i=1, \dots, n-1} t(i)$ (cf., e.g., Csörgő and Horváth (1997)). A smaller γ is usually expected to increase the sensitivity of testing or estimation procedures towards change points in the middle of time series.

In the next subsection we will study estimates under the $d \rightarrow \infty$ asymptotics with respect to the weights (2.5) and (2.6). In particular we will see limitations of (2.5) and propose a suitable extension w^{exact} that has better consistency properties.

2.2.3 Theoretical analysis of weighting schemes

For our analysis we have to impose some (homogeneous) structure on the noise $\{\varepsilon_{i,k}\}$ and on the common factors $\{\zeta_i\}$ in the next two assumptions. Therefore, let

$$S_{i,k}(\varepsilon) = n^{-1/2} \sum_{j=1}^i (\varepsilon_{j,k} - \bar{\varepsilon}_{n,k}) \quad (2.7)$$

⁸If $t(i)$ has a non unique maximum, then counterexamples may be constructed such that $\mathcal{E}(\lambda) = \{\hat{u}\}$ with $\hat{u} \in \arg \max_{i=1, \dots, n-1} t(i)$ is impossible.

⁹When dealing with CUSUM (under $n \rightarrow \infty$ asymptotics), the observations $\{Y_{j,\bullet}\}$ are usually additionally rescaled by the long run covariance matrix.

be the cumulated centered noises in the k -th panel.

Assumption 2.5.

1. The noise $\{\varepsilon_{i,k}\}_{i,k \in \mathbb{N}}$ is centered with finite fourth moments and the variances fulfill $E(\varepsilon_{i,k})^2 = \sigma^2$ for some $0 < \sigma^2 < \infty$ and all i, k .
2. The function

$$V^2(i) = \text{Var}(S_{i,k}(\varepsilon)) / \sigma^2, \quad i = 1, \dots, n-1,$$

is independent of k .

Assumption 2.6. The common factors $\{\zeta_i\}$ are independent of $\{\varepsilon_{i,k}\}$ and are centered with finite fourth moments. Moreover, it holds that, as $d \rightarrow \infty$,

$$\frac{1}{d} \sum_{k=1}^d \gamma_k^2 = o(1). \quad (2.8)$$

Before proceeding further with the theory, we show some specific examples for the function $V^2(i)$ and also discuss some sufficient conditions for part 2 of Assumption 2.5. Clearly, given that part 1 of Assumption 2.5 holds true, a sufficient condition is identical distribution of the panels $\{\varepsilon_{\bullet,k}\}_{k \in \mathbb{N}}$. The following examples will both play important roles in our subsequent analysis.

Example 2.7 (Uncorrelated noise). Assume that part 1 of Assumption 2.5 holds true and that $\{\varepsilon_{i,k}\}_{i \in \mathbb{N}}$ are pairwise uncorrelated for any k . In this situation it holds that

$$V^2(i) = (i/n)(1 - i/n). \quad (2.9)$$

Example 2.8 (Moving average noise). Another interesting case, which satisfies Assumption 2.5, is given by $\{\varepsilon_{i,k}\}_{i,k \in \mathbb{N}}$ where

$$\varepsilon_{i,k} = (\eta_{i,k} + \phi\eta_{i-1,k}) + \theta(\eta_{i,k-1} + \phi\eta_{i-1,k-1}) \quad (2.10)$$

for $i, k \in \mathbb{N}$, i.e. $\{\varepsilon_{i,k}\}_{i,k \in \mathbb{N}}$ are MA(1) in time and across panels. Here, we assume some common parameters $\phi, \theta \in \mathbb{R}$ and centered i.i.d. shocks $\{\eta_{i,k}\}_{i,k \in \mathbb{N}}$ with finite fourth moments and with $E(\eta_{i,k}^2) = \tilde{\sigma}^2$, $0 < \tilde{\sigma}^2 < \infty$. In this case (2.9) extends to

$$V^2(i) = C\alpha(\phi) \left[(i/n)(1 - i/n) \right] - 2C\phi/n \quad (2.11)$$

with

$$\alpha(\phi) = 1 + \phi^2 + 2\phi + 2\phi/n, \quad \sigma^2 = \tilde{\sigma}^2(1 + \phi^2 + \theta^2 + \phi^2\theta^2) \quad (2.12)$$

and with the constant $C = (\tilde{\sigma}/\sigma)^2(1 + \theta^2)$ that is independent of i .

The following deterministic *critical functions* are the cornerstone of our subsequent analysis:

$$C_n(i; u, r) = w^2(i, n) \left[V^2(i)r + H^2(i, u) \right] \quad (2.13)$$

for $0 < r < \infty$ with

$$H(i, u) = \begin{cases} (i/n)(1 - u/n), & i = 1, \dots, u, \\ (u/n)(1 - i/n), & i = u, \dots, n-1, \end{cases}$$

for $u = 1, \dots, n-1$. As before, $u = u_1$ is the change point and $w(i, n)$ are the weights. The parameter

$$r = \left(\frac{1}{n} \frac{\sigma^2}{\Delta_\infty^2} \right), \quad (2.14)$$

for $0 < \sigma^2 < \infty$ and $0 < \Delta_\infty^2 < \infty$, is the normalized *noise to change ratio* where

$$\Delta_\infty^2 = \lim_{d \rightarrow \infty} \left(\frac{1}{d} \sum_{k=1}^d \Delta_k^2 \right) \quad (2.15)$$

is the average change across all panels at the change point u . (Recall from (1.2) and (1.3) that $\Delta_k = \mu_{2,k} - \mu_{1,k}$.) The case $\Delta_\infty^2 = 0$ of a *vanishing change* is excluded but will be addressed in the Remark 2.19 below. For simplicity we will write $C(i; u, n, r) = C_n(i; u, r)$.

The next theorem generalizes¹⁰ Bleakley and Vert (2011a, Lemma 1). Following their approach we show that $C(i; u, n, r)$ is in probability the limit of rescaled $t(i)$ for all $i = 1, \dots, n-1$ (cf. also Proposition 5.1).

Theorem 2.9. *Let Assumptions 2.1, 2.5 and 2.6 be fulfilled. Assume that it holds that, as $d \rightarrow \infty$,*

$$\frac{1}{d^2} \sum_{k,q=1}^d |\text{Cov}(\varepsilon_{j,k} \varepsilon_{h,k}, \varepsilon_{l,q} \varepsilon_{p,q})| = o(1), \quad (2.16)$$

$$\frac{1}{d^2} \sum_{k,q=1}^d \alpha_{k,q} |\text{Cov}(\varepsilon_{j,k}, \varepsilon_{l,q})| = o(1) \quad (2.17)$$

with $\alpha_{k,q} = |\Delta_k \Delta_q| + |\gamma_k \gamma_q|$ and for all $j, h, l, p = 1, \dots, n$. Then, it holds that, as $d \rightarrow \infty$,

$$P \left(\arg \max_{i=1, \dots, n-1} t(i) \subseteq S \right) \rightarrow 1, \quad S = \arg \max_{i=1, \dots, n-1} C(i; u, n, r). \quad (2.18)$$

Theorem 2.9 immediately implies $P(\hat{u}_* \in S) \rightarrow 1$. Note that, in view of Proposition 2.4, the same limiting behaviour holds true for the denoising estimate \hat{u} if the maximum of $C(i; u, n, r)$ is unique - which will be the interesting case in this article.

Remark 2.10. Let Assumptions 2.1, 2.5 and 2.6 hold true. Clearly, conditions (2.16) and (2.17) are fulfilled if $\{\varepsilon_{i,k}\}$ are i.i.d. but deviations, in particular within panels, are also possible. For example if panels $\{\varepsilon_{\bullet,k}\}$ are independent then

$$\sup_{j,h=1, \dots, n, k \geq 1} \text{Var}(\varepsilon_{j,k} \varepsilon_{h,k}) < \infty \quad (2.19)$$

¹⁰Under Gaussianity and independence within panels the results coincide up to a normalizing constant.

is sufficient. Condition (2.19) may be reduced further to $\text{Var}(\varepsilon_{j,1}\varepsilon_{h,1}) < \infty$ for all $j, h = 1, \dots, n$ if we additionally assume identical distribution of $\{\varepsilon_{\bullet,k}\}$. Notice that it is straightforward to check that $\{\varepsilon_{\bullet,k}\}$ from Example 2.8 fulfill conditions (2.16) and (2.17), too.

Perfect estimation and the exact weighting scheme

Given any estimate \tilde{u} for a change point u , we will speak of *perfect estimation*¹¹ if consistent estimation, i.e.

$$P(\tilde{u} = u) \rightarrow 1, \quad (2.20)$$

as $d \rightarrow \infty$, holds true for all possible change points $u = 1, \dots, n-1$ and all possible ratios $0 < r < \infty$.

Theorem 2.9 shows that under mild assumptions the stochastic limits of CUSUM and denoising estimates are described by the deterministic critical function (2.13). This reduces the question of consistency and of perfect estimation to an analytical problem: Given (2.18) it is sufficient for consistency if the following Assumption A1 holds true. Clearly, the same applies to perfect estimation if we require the assumption to hold for all possible change points $u = 1, \dots, n-1$ and all possible ratios $0 < r < \infty$.

Assumption A1. The critical function $C(i; u, n, r)$ has a unique maximum at $i = u$ given a change point at u .

We proceed by studying the existence of weights which ensure perfect estimation for the denoising and the CUSUM estimates. The weights are tacitly assumed not to depend on u .

Theorem 2.11. *Let Assumption 2.1 be fulfilled and assume that V is strictly positive. Only the weights*

$$w^{\text{exact}}(i, n) = \alpha/V(i), \quad \alpha > 0, \quad (2.21)$$

for $i = 1, \dots, n-1$ may fulfill Assumption A1 for all possible change points $u = 1, \dots, n-1$ and all ratios $0 < r < \infty$. For weights other than (2.21) there is some change point u and some ratio $r > 0$ such that the maximum of the critical function $C(i; u, n, r)$ is not at $i = u$.

Since the estimates are independent of any scaling $\alpha > 0$, we may restrict ourselves to the case of $\alpha = 1$ and consider w^{exact} to be unique. Notice that in the setting of Example 2.7 the schemes w^{exact} and w^{standard} coincide and that we already know from Bleakley and Vert (2011a, Theorem 3) that these weights yield perfect estimation for our estimates in the Gaussian i.i.d. setting. However, as follows from Example 2.8, we have generally $w^{\text{exact}} \neq w^{\text{standard}}$ if the noise is dependent in time. Hence, due to Theorem 2.11, weights w^{standard} cannot generally ensure consistency and perfect estimation in such cases¹². Notice that the weightings w^{standard} and w^{exact} might differ fundamentally.

¹¹Note that perfect estimation was previously defined in Torgovitski (2015) in terms of the limiting critical function C .

¹²CUSUM (and denoising) estimates are not consistent if the (unique) maximum of $C(i; u, n, r)$ is not at $i = u$.

The former is strictly convex but the latter is even strictly concave in case of (2.11) for $\alpha(\phi) < 0$.

A consequence of the next theorem, under the assumptions of Theorem 2.9, is that covariance-based weights w^{exact} ensure perfect estimation under additional (but again) quite mild assumptions on $V(i)$. Therefore, we need to find conditions such that $H(i, u)/V(i)$ has a (unique) maximum at $i = u$ for any $u = 1, \dots, n - 1$.

Theorem 2.12. *Let Assumption 2.1 be fulfilled. Assume that we use the weights w^{exact} , that V is strictly positive and that $V(i) = V(n-i)$ holds true for all i . Then Assumption A1 is fulfilled for all possible change points $u = 1, \dots, n - 1$ and all ratios $0 < r < \infty$ if and only if*

$$V(i)/i > V(u)/u, \quad i, u = 1, \dots, n - 1, \quad i < u \quad (2.22)$$

holds true. This condition is equivalent to $V(i)/i$ being strictly decreasing.

Note that the symmetry $V(i) = V(n - i)$ for all $i = 1, \dots, n - 1$ is implied by the symmetry of covariances

$$\text{Cov}((\varepsilon_{1,1}, \dots, \varepsilon_{n,1})^T) = \text{Cov}((\varepsilon_{n,1}, \dots, \varepsilon_{1,1})^T) \quad (2.23)$$

and a sufficient condition for (2.23) is weak stationarity of $\{\varepsilon_{i,1}\}$. The next lemma provides, based on concavity, a condition for (2.22) which is sometimes easier to verify than monotonicity of $V(i)/i$ and that will also be used in Remark 2.14. It is not clear how to state a comparable condition under convexity.

Lemma 2.13. *Assume that $V(i)$ is strictly concave¹³, strictly positive and that $n \geq 3$. Then (2.22) already holds true if*

$$V(1) > V(u)/u \quad (2.24)$$

holds true for all $u = 2, \dots, n - 1$.

Remark 2.14. The function $V(i)$ from Example 2.8 fulfills (2.22) with any moving average parameters $\phi, \theta \in \mathbb{R}$ and is always strictly positive for $n \geq 3$.

Now, a combination of Theorems 2.9 and 2.12 together with Remarks 2.10 and 2.14 yield the following corollary.

Corollary 2.15. *Consider the moving average noise from Example 2.8 and let Assumptions 2.1 and 2.6 hold true. Using w^{exact} we have perfect estimation for denoising and CUSUM estimates for any moving average parameters $\phi, \theta \in \mathbb{R}$.*

Consistent estimation and the weighted weighting scheme

We turn to the analysis of w^{weighted} and, analogously to Peřtov and Peřta (2015, Theorem 3), our aim is to identify noise to change ratios r for which consistent estimation (2.20) does or does not hold true for the denoising and the CUSUM estimates. We restrict the consideration to panels which fulfill all assumptions of Theorem 2.9 with $V^2(i) = (i/n)(1 - i/n)$ from Example 2.7 (cf. Remark 2.10). We expect more restrictive

¹³As a discrete function linearly interpolated on the interval $[1, n - 1]$.

conditions on ratios r for changes u closer to the edges of the samples, and vice versa less restrictive conditions if u is more centered. This expectation is confirmed by the next theorem where the minimum is taken over smaller sets in (2.25) if u is closer to $n/2$.

Theorem 2.16. *Let Assumption 2.1 be fulfilled. Assume $V(i)$ as in (2.9) and assume that we use w^{weighted} with $\gamma \in [0, 1/2)$. Define $R(u, i)$ as in (5.16) below and set*

$$\mathcal{R} := \min_{n/2 \leq i < u^*} R(u^*, i) > 0 \quad (2.25)$$

with

$$u^* = \begin{cases} n - u, & u = 1, \dots, \lfloor n/2 \rfloor, \\ u, & u = \lceil n/2 \rceil, \dots, n - 1 \end{cases}$$

and where $\min \emptyset = \infty$. Then Assumption A1 holds true for $0 \leq r < \mathcal{R}$ and does not hold true for $\mathcal{R} < r$. In the latter case the maximum of the critical function $C(i; u, n, r)$ is not at $i = u$.

The bound in (2.25) can be evaluated numerically but to gain more insight into the influence and the interaction of parameters, it is desirable to get explicit representations and approximations of this expression. We will provide such approximations where we first let $d \rightarrow \infty$ and then consider $n \rightarrow \infty$. Therefore, we have to introduce a *boundary function*

$$B(\gamma) = \begin{cases} \frac{4\gamma^2 + 6\gamma^{3/2} - 3\gamma^{1/2} - 1}{8\gamma^2 + 8\gamma^{3/2} - 4\gamma^{1/2} - 2\gamma - 1}, & \gamma \in [0, 1/2), \\ 2^{-1/2}, & \gamma = 1/2. \end{cases} \quad (2.26)$$

This function $B(\gamma)$ is monotonously decreasing, continuous and $B(\gamma) \geq 2^{-1/2}$ holds true which can be seen as follows. It holds that $\partial_\gamma(B(\gamma^2)) = -(1 + 2\gamma)^{-2}$ which in turn implies that $\partial_\gamma B(\gamma) < 0$ holds true for any $\gamma \in (0, 1/2)$. Applying l'Hôpital's rule to $B(\gamma^2)$ we get the continuity at $\gamma = 1/2$.

Theorem 2.17. *Let the assumptions of Theorem 2.16 hold true, set $B(\gamma)$ as in (2.26), set $s = u^*/n$ and let $\mathcal{R} = \mathcal{R}(s, \gamma)$ be as in (2.25).*

1. *If $1/2 + 1/n < s \leq B(\gamma)$, then $\mathcal{R}(s, \gamma)$ equals to the unique solution of*

$$C(u^* - 1; u^*, n, r) = C(u^*; u^*, n, r) \quad (2.27)$$

and for $u^ = \lfloor n\zeta \rfloor$ with any $1/2 < \zeta \leq B(\gamma)$ it holds that*

$$\lim_{n \rightarrow \infty} \mathcal{R}(s, \gamma) = \zeta(a - \zeta)(b - \zeta)(\zeta - c)^{-1} \quad (2.28)$$

with $a = 1$, $b = (\gamma - 1)/(2\gamma - 1)$, $c = 1/2$.

2. *If $B(\gamma) + 2/n < s < 1$, then the unique solution to (2.27) is larger than $\mathcal{R}(s, \gamma)$.*

In (2.28) the quantity $\mathcal{R}(s, \gamma)$, i.e. the range for r that ensures consistency, becomes larger for parameters γ and ζ close to $1/2$, which confirms our intuition. Note that Theorem 2.17 is consistent with Bleakley and Vert (2011a, Theorem 2) since, as $n \rightarrow \infty$, it holds

$$\mathcal{R}(s, 0) = \zeta(1 - \zeta)^2(\zeta - 1/2)^{-1} + o(1)$$

and $B(0) = 1$.

The weighting with $\gamma = 1/4$ can be seen as a compromise between w^{standard} and w^{simple} . For this particular γ we are able to compute $\lim_{n \rightarrow \infty} \mathcal{R}(s, 1/4)$ for any $\zeta \in (1/2, 1)$. It would be interesting to know if such a formula could be computed for the remaining parameters $\gamma \in (0, 1/4) \cup (1/4, 1/2)$ as well.

Proposition 2.18. *Let the assumptions of Theorem 2.17 hold true, set $\gamma = 1/4$ and $B(\gamma)$ as in (2.26), i.e. $B(1/4) = 3/4$. Then it holds for $\mathcal{R} = \mathcal{R}(s, 1/4)$ that*

$$\lim_{n \rightarrow \infty} \mathcal{R}(s, 1/4) = \begin{cases} \frac{\zeta(1-\zeta)(3/2-\zeta)}{(\zeta-1/2)}, & \zeta \in (1/2, 3/4], \\ \frac{(\zeta-1)^2(\zeta(\zeta+1)+(\zeta/2-1)(1-\zeta)^{1/2}-1)}{\zeta(\zeta-1)+(\zeta/2)(1-\zeta)^{1/2}}, & \zeta \in (3/4, 1) \end{cases} \quad (2.29)$$

This function is continuously differentiable for $\zeta \in (1/2, 1)$. The second derivative does however not exist for $\zeta = 3/4$.

Spurious estimation for vanishing change points

We close this subsection by a short remark on spurious estimation for $\Delta_\infty = 0$, i.e. a probably common change that vanishes asymptotically. We also include the case $\Delta_k = 0$ for $k \geq 1$.

Remark 2.19. Let all assumptions of Theorem 2.9 hold true but with $\Delta_\infty = 0$. Following the proof of Theorem 2.9 it is clear that (2.18) also holds true in this situation with $C(i; u, n, r) := w^2(i, n)V^2(i)$. In case of Example 2.7 and for w^{weighted} this yields

$$\arg \max_{i=1, \dots, n-1} C(i; u, n, r) = \begin{cases} \{\lfloor n/2 \rfloor, \lceil n/2 \rceil\}, & \gamma \in [0, 1/2), \\ \{1, \dots, n-1\}, & \gamma = 1/2, \end{cases}$$

i.e. estimation of spurious changes. In case of Example 2.8 and using w^{exact} it also always holds that $\arg \max_{i=1, \dots, n-1} C(i; u, n, r) = \{1, \dots, n-1\}$ since $C(i; u, n, r)$ is constant in i .

2.2.4 Estimation of the exact weighting scheme

In this subsection we discuss estimates of $w^{\text{exact}}(i, n)$, or equivalently of $V(i)$, under the assumptions that V is strictly positive, that the sequences $\{\varepsilon_{j,k}\}_{k \in \mathbb{N}}$ and $\{\varepsilon_{j,k} \varepsilon_{l,k}\}_{k \in \mathbb{N}}$ fulfill the weak law of large numbers for any $1 \leq j, l \leq n$ and that Assumptions 2.1, 2.5 and 2.6 hold true. Notice that all conditions on the noise $\{\varepsilon_{j,k}\}$ are clearly fulfilled for our moving average Example 2.8.

The partial sums (2.7) can be rewritten as

$$S_{i,k}(\varepsilon) = \sum_{j=1}^n a_{i,j} \varepsilon_{j,k} \quad (2.30)$$

with

$$a_{i,j} = \frac{1}{n^{1/2}} \begin{cases} 1 - i/n, & j = 1, \dots, i, \\ -i/n, & j = i + 1, \dots, n \end{cases} \quad (2.31)$$

and it is straightforward to check that

$$V^2(i) = f_i(\Sigma)/\sigma^2, \quad (2.32)$$

where $\Sigma = \text{Cov}(Y_{\bullet,1})$, with functions $f_i(\Sigma) = a_{i,\bullet} \Sigma a_{i,\bullet}^T$ for $i = 1, \dots, n-1$.

A natural estimate for Σ is given via $\hat{\Sigma}$ with

$$\hat{\Sigma}_{j,k} = \frac{1}{d-1} \sum_{p=1}^d (Y_{j,p} - \bar{Y}_{j,d})(Y_{k,p} - \bar{Y}_{k,d}), \quad (2.33)$$

where $1 \leq j, k \leq n$ and $d > 1$. This estimate $\hat{\Sigma}_{j,k}$ is consistent¹⁴, as $d \rightarrow \infty$, e.g. if we assume additionally that $\mathcal{M}_{i,k} = c_i$ always holds true for some $c_i \in \mathbb{R}$ and all $k \geq 1$, i.e. that at all time points the means across all panels are the same.

Now, $V^2(i)$ simply may be estimated by $\hat{V}^2(i) = f_i(\hat{\Sigma})/\sigma^2$ and the corresponding estimate for the weights will be denoted by \hat{w}^{exact} . Here, $\bar{Y}_{j,d}$ is the mean over all panels at time point j . Recall also, that our change point estimates do not depend on the scaling of the weights $w^{\text{exact}}(i, n)$. Thus, without loss of generality we may assume here that $\sigma^2 = 1$ and technically we do not have to estimate this parameter in (2.32).

Remark 2.20. Reasonable estimates for the exact weights have to be strictly positive which corresponds to positiveness of $f_i(\hat{\Sigma})$. This is ensured asymptotically with probability tending to 1, as $d \rightarrow \infty$, because the estimate $\hat{\Sigma}$ is consistent and because of our usual assumption $V(i) > 0$. (Clearly, a sufficient condition for finite d would be the positive definiteness of the estimate $\hat{\Sigma}$.)

Banded estimation based on a training period

Now, we aim to increase the precision of the estimate for Σ by averaging and by a banded covariance approach. To do so we have to impose some structural assumptions: We assume weak stationarity for $\{\varepsilon_{j,k}\}_{j \in \mathbb{N}}$ to hold in time, i.e. within each panel and additionally, we assume to have a *training period* between n_1 and n_2 , $1 \leq n_1 < n_2 \leq n$ where the above assumptions of $\mathcal{M}_{i,k} = c_i$ hold true for $i = n_1, \dots, n_2$ and $k \geq 1$. Within this training period we can compute consistent estimates $\hat{\Sigma}'_{j,k} \approx \text{Cov}(\varepsilon_{j,1}, \varepsilon_{k,1})$,

¹⁴The calculations are straightforward and the contribution of common factors vanishes asymptotically due to Assumption 2.6.

according to (2.33), for $j, k = n_1, \dots, n_2$ and then may average these estimates as follows

$$\xi_r = \frac{1}{n_2 - n_1 - r + 1} \sum_{j=n_1}^{n_2-r} \hat{\Sigma}'_{j,j+r} \quad \left(\approx \text{Cov}(\varepsilon_{1,1}, \varepsilon_{1+r,1}) \right) \quad (2.34)$$

for $r = 0, \dots, n_2 - n_1$ to gain more stability. (These estimates are consistent as well.) Finally, the desired banded estimate $\hat{\Sigma}$ is obtained via

$$\hat{\Sigma}_{j,j+r} = \hat{\Sigma}_{j+r,j} = \begin{cases} \xi_r, & r = 0, \dots, \min\{h, n - j\}, \\ 0, & r = \min\{h, n - j\} + 1, \dots, n \end{cases} \quad (2.35)$$

for $j = 1, \dots, n$ and with some banding parameter $h \in \{0, \dots, n_2 - n_1\}$ to be chosen. Clearly, the corresponding estimate $\hat{V}^2(i) = f_i(\hat{\Sigma})/\sigma^2$ is consistent, as $d \rightarrow \infty$, for panels that are MA(q) in time if $q \leq h \leq n_2 - n_1$ holds true. Heuristically, it yields also a reasonable approximation for $\hat{V}^2(i) = f_i(\hat{\Sigma})/\sigma^2$ in case of stationary and weak dependent panels, again in the time-domain, whenever the covariances $\text{Cov}(\varepsilon_{1,1}, \varepsilon_{1+r,1})$ for $r > h$ are negligible. The corresponding estimate for the weights will be denoted by $\hat{w}^{\text{exact-banded}}$. For a data-driven selection of the banding parameter h one could use the approach of [Bickel and Levina \(2008, Section 5 and disp. \(24\)\)](#) (cf., e.g., also [Wu and Pourahmadi \(2009\)](#)). In their simulations for MA(1) covariance structure, the banding parameter is always chosen correctly, i.e. $h = 1$.

The assumptions stated above are restrictive and may be questionable in applications. In the following we (informally) discuss estimates for more complex situations when $\mathcal{M}_{i,k} = c_i$ for $k \geq 1$ does not hold true within a reasonable subsample. Again, we need to assume a training period between n_1 and n_2 , $1 \leq n_1 < n_2 \leq n$, such that either $n_2 \leq u$ or $n_1 > u$ holds true, i.e. that a common change does not occur in this subsample. Further, we need to assume stationarity and the weak law of large numbers to hold in time for $\{\varepsilon_{j,k}\}_{k \in \mathbb{N}}$, i.e. within each panel. Now, in the first step, we center each panel based on means computed within the training period and for each panel separately, i.e.

$$Y'_{i,k} := Y_{i,k} - \frac{1}{n_2 - n_1 + 1} \sum_{j=n_1}^{n_2} Y_{j,k}$$

for $i = 1, \dots, n$ and $k = 1, \dots, d$. In the next step (as before) we compute only estimates $\hat{\Sigma}'_{j,k}$ for $n_1 \leq j \leq k \leq n_2$ but now based on the centered panels $\{Y'_{j,k}\}$. Proceeding as under (2.34) and (2.35) we obtain $\hat{V}^2(i) = f_i(\hat{\Sigma})/\sigma^2$ and the corresponding weights $\hat{w}^{\text{exact-banded}}_{\text{centered}}$. Those are heuristically reasonable for large n , d and a relatively large training period which is backed up by our simulations. However, to formalize this one would need to consider asymptotics $n, d \rightarrow \infty$ with $|n_2 - n_1| \rightarrow \infty$ which is not in the scope of this paper.

3 Simulations

For our simulations within the single change point scenario we have implemented the estimates (2.4)¹⁵ in MATLAB. For demonstration purposes an application with a graphical

¹⁵We choose \hat{u}_* as the smallest element in $\arg \max t(i)$.

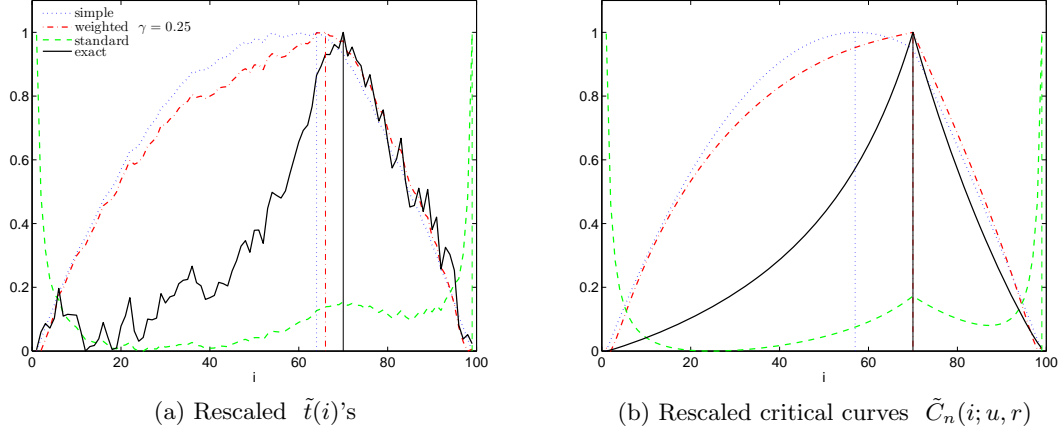


Figure 2: Estimation of the change point at $u = 70$ with different weighting schemes. The parameters are $\sigma^2 = 9$, $\phi = -3$, $\theta = 1$ and $d = 10000$.

user interface can be obtained from the author or from www.mi.uni-koeln.de/~ltorgovi. For the simulations within the multiple change point scenario we work with the MATLAB “GFLseg”-package of [Bleakley and Vert \(2011b\)](#). Notice that the denoising approach may be interpreted as a group fused LASSO (least absolute shrinkage and selection operator) (cf. Section 5) and that the corresponding group fused LARS (least angle regression) method yields a fast approximation to the LASSO solution (cf. [Bleakley and Vert \(2011a\)](#)). In particular, we use the LASSO and LARS methods that are implemented in `glassoK.m` or in `glars.m` respectively.

We proceed by considering panels $\{Y_{i,k}\}$ with the following parameters: The noise $\{\varepsilon_{i,k}\}$ is moving average from Example 2.8 based on independent Gaussian innovations $\{\eta_{i,k}\}$. The common factors $\{\zeta_i\}$ are chosen to be independently uniformly distributed, centered and with the same variance as the noise $\{\eta_{i,k}\}$. Unless stated otherwise, the factor loadings are set to $\gamma_k = k^{-1/2}$ and the length of panels is $n = 100$. For simplicity, we choose common changes with $\mu_{1,k} = 0$ and $\mu_{2,k} = 1$ for all k , i.e. $\Delta_k = 1$ and therefore $\Delta_\infty = 1$. For $\hat{w}^{\text{exact-banded}}$ and for $\hat{w}^{\text{exact-banded}}_{\text{centered}}$ the training period is chosen as $n_1 = 1$ and $n_2 = 20$ with a bandwidth $h = 2$. (The influence of the misspecification of h is rather mild in our settings.)

All Monte Carlo simulations will be based on 100 repetitions. Notice that w^{standard} replaces \hat{w} , whenever the corresponding estimate $\hat{V} = 1/\hat{w}$ has at least one non-positive entry (cf. Remark 2.20).

3.1 Segmentation under dependence

Figure 2 shows one realization of $\tilde{t}(i)$ ’s from (2.4) for different weighting schemes¹⁶ and the corresponding critical curves $\tilde{C}_n(i; u, r)$ from (2.13), which are in probability the limits of the $\tilde{t}(i)$ ’s as $d \rightarrow \infty$. The vertical lines indicate the locations of maxima for the respective weightings, i.e. the positions of estimates \hat{u}_\star and \hat{u} . We see that w^{exact} provides a correct estimation of $u = 70$ whereas w^{standard} does not. For w^{simple} we estimate a more centered change point and w^{standard} estimates a completely wrong location at the

¹⁶For the sake of comparison we shift and rescale the curves by the transformation: $f(i) \mapsto \tilde{f}(i) = [f(i) - \min_{1 \leq j < n} f(j)] / [\max_{1 \leq j < n} f(j) - \min_{1 \leq j < n} f(j)]$.

right border.

Figures 3 and 4 demonstrate the performance of the estimates with respect to different weighting schemes. They compare the accuracy $P(\hat{u}_\star = u)$, the means and the standard deviation of the estimate \hat{u}_\star . We ran simulations with $\theta = 1$ and $\tilde{\sigma}^2 = 9$ and considered a range of parameters: $\phi = -2, -1, -0.5, 0, 1$ and $d = 50, 150, 250, 200, \dots, 1950, 2000$.

The figures show that the change point estimate based on the (estimated) exact weighting scheme w^{exact} outperforms w^{standard} . The former estimates all changes correctly for arbitrary ϕ if we consider a sufficiently large number of panels d . The exact scheme is less biased and also has overall less variation for negative ϕ . Furthermore, we see for chosen parameters that the distortion due to estimation with $\hat{w}_{\text{centered}}^{\text{exact-banded}}$ is rather mild¹⁷. However, it might be strong for other parameters e.g. if the training period is too small. Notice that change point estimation with \hat{w}^{exact} may (surprisingly) perform even better than with w^{exact} (cf. Figure 4).

3.2 Segmentation under varying change point locations

So far, we assumed a fixed change point at some time point u across all panels. It seems more realistic to allow for some minor fluctuations (around time point u) of change points in different panels. Therefore, Bleakley and Vert (2011a) studied the behaviour of their procedure under randomized change points theoretically and empirically. They considered changes across panels $k = 1, \dots, d$ that are located at random change points $u + U_k \in \{1, \dots, n - 1\}$, where $\{U_k\}$ are some i.i.d. random variables describing the fluctuations¹⁸. In (Bleakley and Vert, 2011a, cf. Theorem 4 and Figure 3) they showed under appropriate assumptions that the standard weighting works also well in this setting in the sense that the probability $P(\hat{u}_\star \in u + S)$ tends to 1 as $d \rightarrow \infty$, where S is the support of P_{U_1} . We do not develop the theoretical analogue but show in Figure 5 empirically that, as should be expected, the exact weighting tends to be beneficial under dependence. For this simulation we stick to the panels and simulation parameters of Subsection 3.1 with $\tilde{\sigma}^2 = 9$ and $\theta = 1$. As in Bleakley and Vert (2011a), we assume $P(U_k = \pm 2) = 0.5$ and we use the term *accuracy* now for $P(\hat{u}_\star \in u + S)$.

3.3 Segmentation in the multiple change point scenario

In this subsection we assume multiple change points and compare the standard weighting scheme with the exact one using the denoising approach. First, we discuss *epidemic* changes, i.e. we have two change points u_1 and u_2 where the means are temporarily shifted after u_1 but return to their former states after u_2 .

We performed various simulations for different change point locations u_1 and u_2 , moving average parameters ϕ and θ and for different variances $\tilde{\sigma}^2$ where we restricted our considerations to a simplified setting with the same magnitude of changes in all panels. We observed that the exact weighting tends to be beneficial in the sense that the overall picture improves. This is demonstrated in Figure 6: With the exact weighting the accuracy for $\phi < -1$ increases considerably whereas for $\phi \geq -1$ it only decreases slightly. The curves are obtained using the fast LARS method but the group fused LASSO yields similar results.

¹⁷The results for $\hat{w}^{\text{exact-banded}}$ are close to $\hat{w}_{\text{centered}}^{\text{exact-banded}}$ in this particular setting and are therefore omitted.

¹⁸ $\{U_k\}$ are assumed to be also independent of $\{\varepsilon_{i,k}, \zeta_i; i, k \in \mathbb{N}\}$.

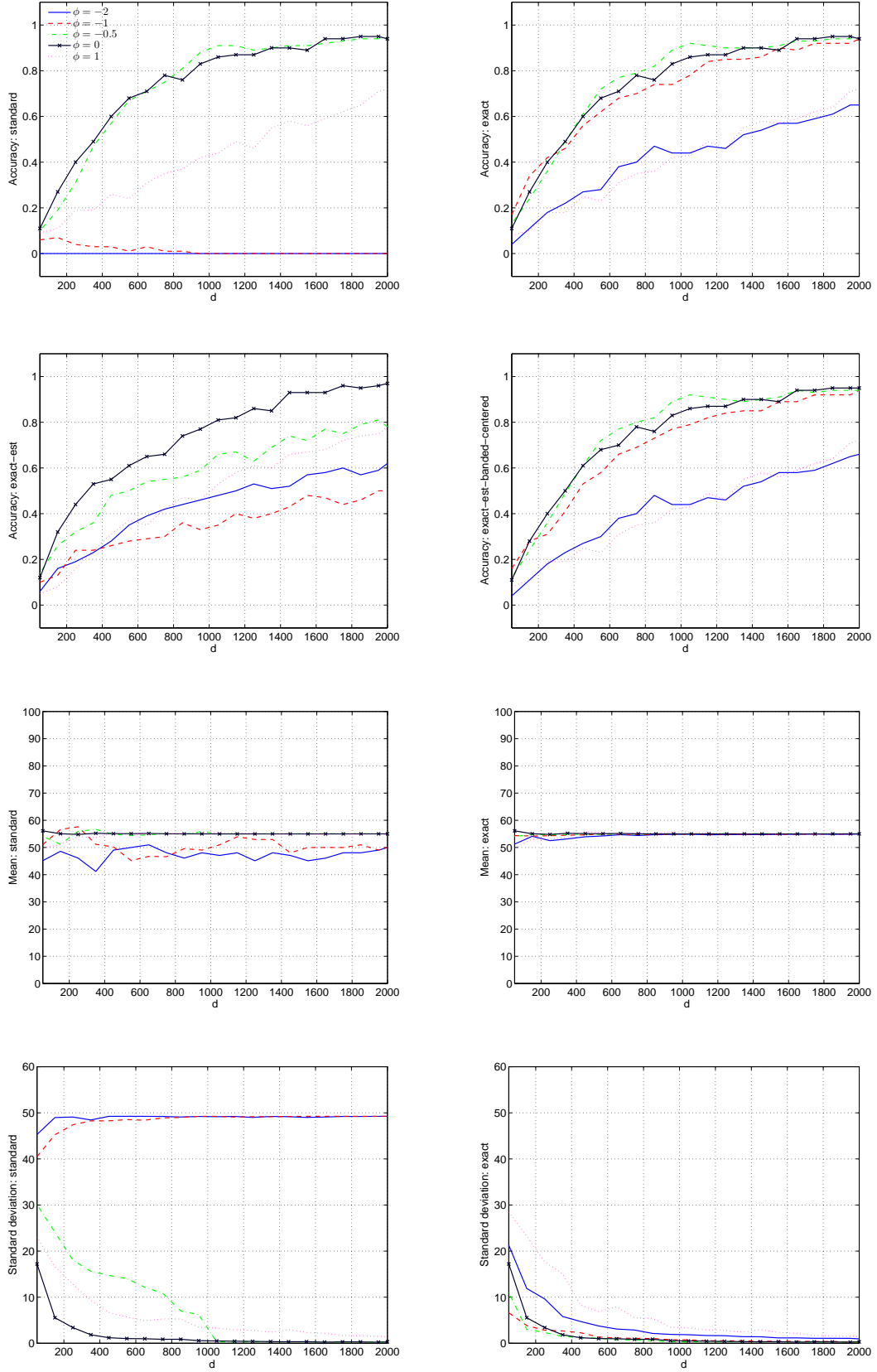


Figure 3: Accuracy, mean and standard deviation for $u = 55$. \hat{w}^{exact} is denoted by “exact-est” and $\hat{w}_{\text{centered}}^{\text{exact-banded}}$ is denoted by “exact-est-banded-centered”.

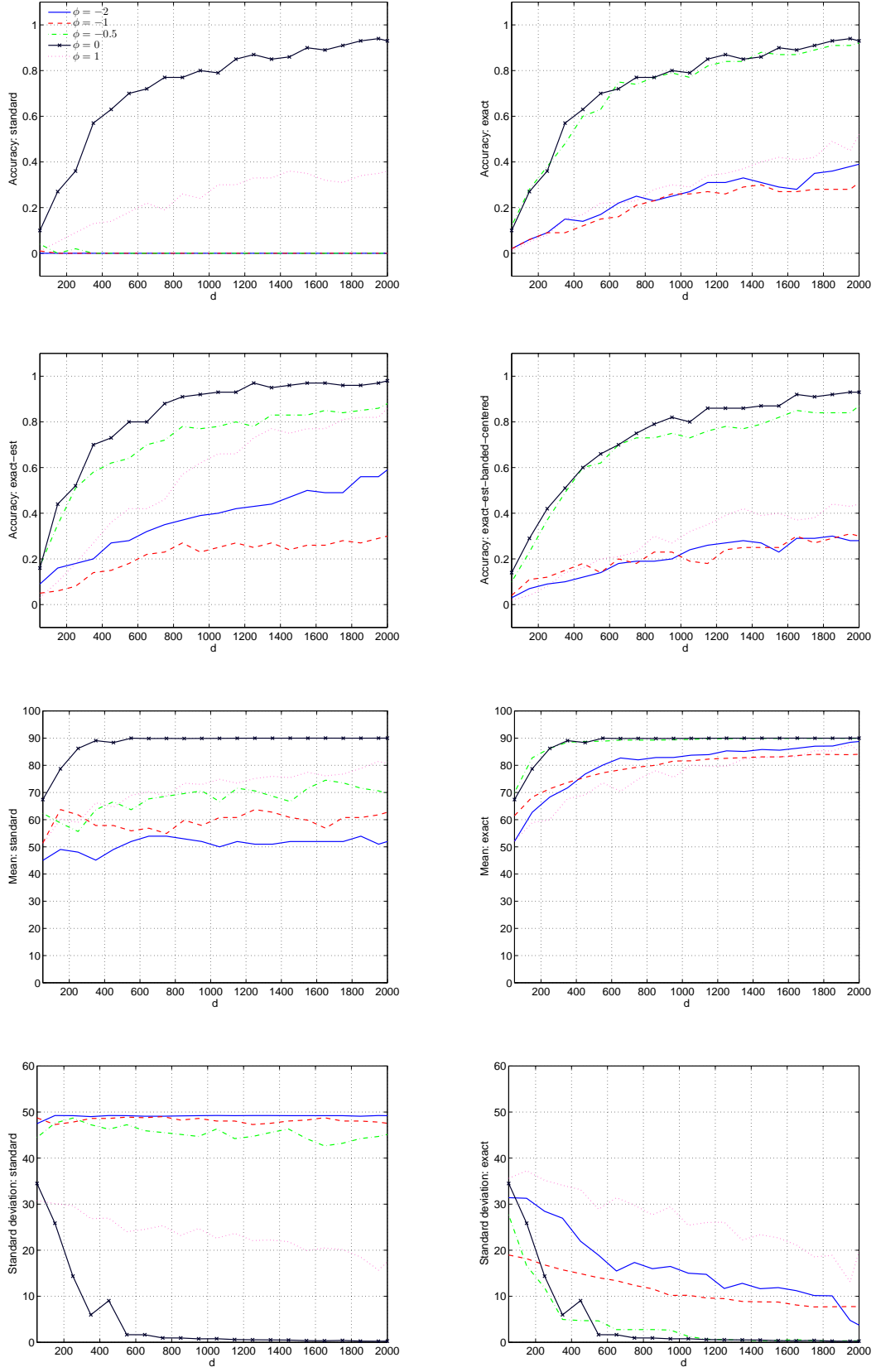


Figure 4: Accuracy, mean and standard deviation for $u = 90$

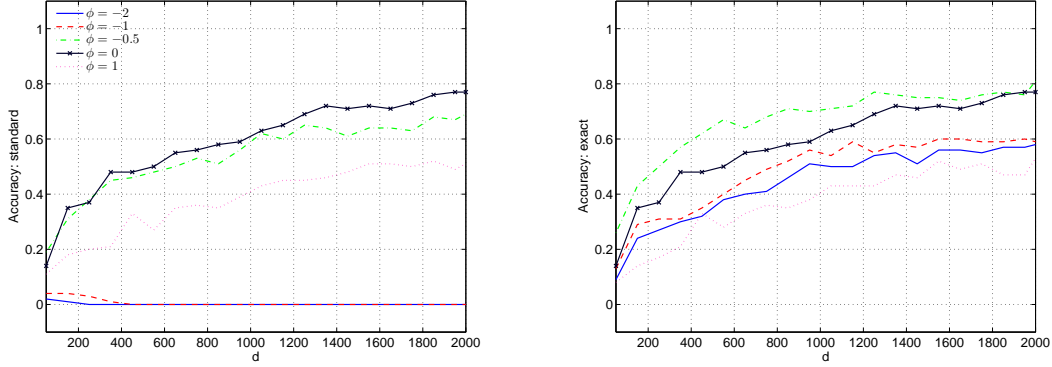


Figure 5: Random change locations $u + U_k$ that fluctuate around $u = 70$. The means are $m_{j,k} = 0$ for $j \leq u + U_k$ and $m_{j,k} = 1$ for $j > u + U_k$ for all k .

In general multiple change point settings the situation is less clear than in the single change point or epidemic settings and the behaviour is rather erratic: The results strongly depend on the location and on the magnitude of the changes as well as on the moving average parameters - it is possible to find settings where the exact weighting scheme outperforms the standard one but also vice versa.

3.4 Post processing estimated exact weights using regression

Here, our interest is again in the single change point scenario using \hat{u}_\star with the simple weighting estimate $\hat{w}(i) = \hat{w}^{\text{exact}}(i, n)$ which is based on (2.33). We would like to mention a possible consistent modification which tends to be beneficial in situations described below and that may serve as a motivation for further research. In the following we assume that w^{exact} is strictly convex since the strictly concave case can be treated analogously.

Based on the results and the corresponding proofs of Section 2 it seems reasonable to expect that, if w^{exact} is strictly convex and smooth, which is the case for panels based on Example 2.7 or on Example 2.8 (with e.g. $\phi \geq 0$), then modifications of \hat{w} that are strictly convex, and therefore also less oscillating, should increase the precision of the resulting change point estimate. Obviously, the estimate \hat{w} is usually not strictly convex due to the fluctuations around the strictly convex (discrete) function w . To obtain a smoother convex estimate $\hat{w}^{\text{exact-reg}} = \tilde{w}$, one may post-process the weights \hat{w} using the well known least squares convex regression. The basic principle is that, given a regression model

$$\hat{w}(i) = w(i) + \varepsilon_i$$

with a strictly convex function $w(i)$ and some centered noise sequence $\{\varepsilon_i\}$ for $i = 1, \dots, n-1$, we solve

$$\text{Minimize}_{\tilde{w}(i), g_i} \sum_{i=1}^{n-1} [\tilde{w}(i) - \hat{w}(i)]^2,$$

under the convexity restrictions

$$\tilde{w}(j) \geq \tilde{w}(i) + g_i(j - i) \tag{3.1}$$

where (3.1) holds true for $i, j = 1, \dots, n-1$ (cf., e.g., Boyd and Vandenberghe (2004) and Hannah and Dunson (2013)). Clearly, in our situation, these estimates $\tilde{w}(i)$ remain consistent for $d \rightarrow \infty$ if the underlying original estimates $\hat{w}(i)$ were consistent and therefore strictly convex with probability tending to 1, as $d \rightarrow \infty$.

In our simulations¹⁹ with moving average panels we observe that the weighting schemes $\hat{w}^{\text{exact-reg}}$ and w^{exact} tend to yield similar estimation results for \hat{u}_* . They outperform \hat{w}^{exact} for smaller variances $\tilde{\sigma}^2$, smaller panel numbers d and for parameters ϕ that are closer to -1 (cf. Figure 7). This effect is stronger for change points u closer to $n/2$ and weaker for u closer to n . Notice that the estimate \hat{w}^{exact} outperforms the “true” w^{exact} for larger variances and larger panel numbers d . This goes in line with the observations of Subsection 3.1.

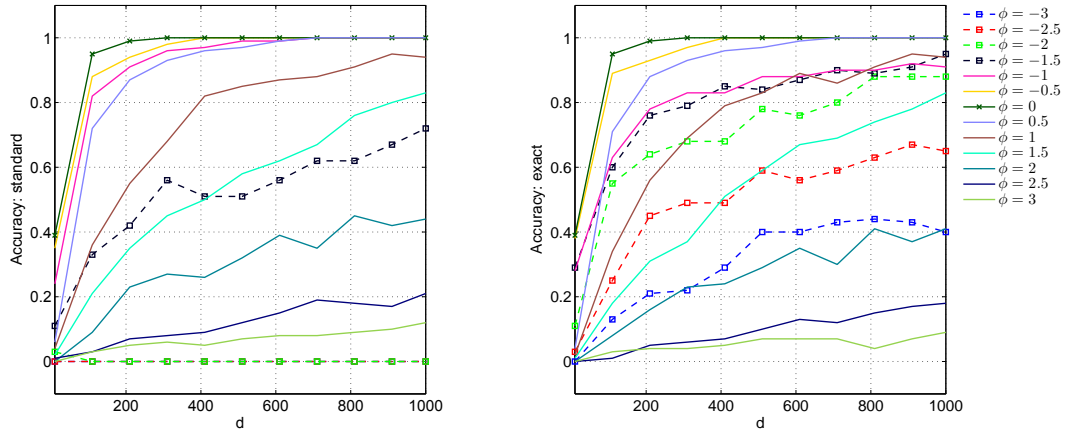


Figure 6: Epidemic change with change points at $\{55, 80\}$ and a jump of size $+1$ in all panels. We consider $\tilde{\sigma}^2 = 1$ with $\theta = 0$ and we do not take common factors into account, i.e. we set $\gamma_k = 0$. *Accuracy* denotes now the probability that all change points are estimated correctly. Notice that the curves for $\phi = -2, -3.5$ are both zero in the left figure.

¹⁹The computation of the regression weights is carried out using the MATLAB software “CVX: A system for disciplined convex programming” (see <http://cvxr.com/cvx/>).

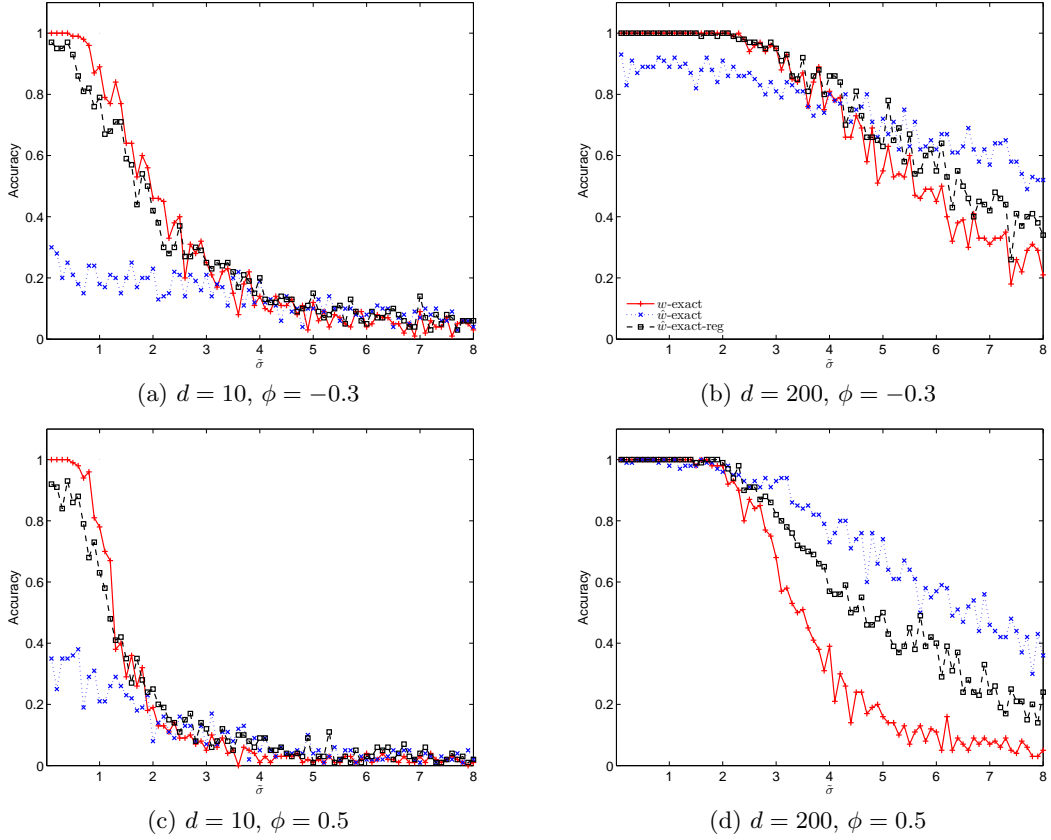


Figure 7: We consider panels of length $n = 50$ with a change point at $u = 37$ and with $\theta = 0$. We do not take common factors into account, i.e. we set $\gamma_k = 0$.

4 Conclusion

In this article we showed the connection of the total variation denoising approach of [Bleakley and Vert \(2011a\)](#) to the classical weighted CUSUM estimates. We generalized the consistency results of [Bleakley and Vert \(2011a\)](#) to panels of time series in the fixed n and $d \rightarrow \infty$ setting under mild assumptions and studied consistency properties with respect to a well-known class of weighting schemes. Doing so, we also defined the criterion of *perfect estimation*, which is fulfilled in the independent setting if one uses the standard weighting, and showed that generally only a suitable covariance-dependent modification of these weights ensures this criterion. Thus, corresponding estimates outperform the standard weighting in various situations. We discussed appropriate estimation of these new weights and confirmed our results in a detailed simulation study. Moreover, we discussed the implications and possible advantages for the multiple change points and the random change point scenarios as well.

5 Proofs

We start by proving Proposition 2.4, which requires some preliminary considerations and some key facts from Bleakley and Vert (2011a). For theoretical investigations and also for practical purposes the authors pick up the idea of Harchaoui and Lévy-Leduc (2008) and reformulate the minimization Problem (2.1) as a *group fused LASSO*.

Using the weights $w(i, n)$, they introduce a fixed design matrix $D \in \mathbb{R}^{n \times (n-1)}$ with

$$D_{i,j} = \begin{cases} w(j, n), & i > j, \\ 0, & \text{else} \end{cases}$$

and set

$$\beta_{i,\bullet} = \frac{U_{i+1,\bullet} - U_{i,\bullet}}{w(i, n)}, \quad (5.1)$$

which can be compactly rewritten as $U = \mathbf{1}U_{1,\bullet} + D\beta$ where $\mathbf{1} = [1, \dots, 1]^T$. Thereby, the original problem (2.1) transforms to

$$\text{Minimize}_{\beta \in \mathbb{R}^{(n-1) \times d}} \frac{1}{2} \|\bar{Y} - \bar{D}\beta\|_F^2 + \lambda \sum_{i=1}^{n-1} \|\beta_{i,\bullet}\|_2 \quad (5.2)$$

with the same $\lambda \geq 0$, where \bar{Y} and \bar{D} are the column-wise centered matrices Y and D , respectively. Let $\hat{\beta}(\lambda)$ denote the solution of (5.2). The solution \hat{U} of (2.1) can be recovered via $\hat{U} = \mathbf{1}\hat{\gamma} + D\hat{\beta}$ with $\hat{\gamma} = \mathbf{1}^T(Y - D\hat{\beta})/n$. The indices of non-zero rows of matrix $\hat{\beta}$ correspond to the change point set \mathcal{E} in (2.3) via (5.1) and it holds that $\mathcal{E}(\lambda) = \{u \mid \hat{\beta}_{u,\bullet}(\lambda) \neq 0\}$.

The crucial observation for further theoretical analysis is that β minimizes (5.2), for any fixed λ , if it fulfills the necessary and sufficient Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned} G_i &= \lambda B_i & \forall \beta_{i,\bullet} \neq 0, \\ \|G_i\| &\leq \lambda & \forall \beta_{i,\bullet} = 0, \end{aligned} \quad (5.3)$$

for all $i = 1, \dots, n-1$ with vectors $G_i = \bar{D}_{\bullet,i}^T (\bar{Y} - \bar{D}\beta)$ and $B_i = \beta_{i,\bullet} / \|\beta_{i,\bullet}\|$.

The next proposition formalizes the selection of the regularization parameter λ which was already informally described in the Subsection 2.2.

Proposition 5.1. *Consider the random matrix $\hat{c} = \bar{D}^T \bar{Y}$, set $t_i = \|\hat{c}_{i,\bullet}\|$ for $i = 1, \dots, n-1$. Assume $t_{i_1} \leq t_{i_2} \leq \dots \leq t_{i_{n-1}}$ with $i_k \neq i_r$ for $k \neq r$ and set $M = i_{n-1}$, $m = i_{n-2}$. Under Assumption 2.1 it holds that:*

1. If $\lambda \geq t_M$ then $\beta = 0$ solves the KKT system (5.3).
2. If $t_m < \lambda < t_M$ then a random λ_{\min} exist such that for any $\lambda_{\min} < \lambda < t_M$ the β with rows

$$\beta_{i,\bullet} = \alpha \begin{cases} \hat{c}_{M,\bullet}, & i = M, \\ 0, & i \neq M, \end{cases} \quad (5.4)$$

and $\alpha = (t_M - \lambda) / (\bar{D}_{\bullet,M}^T \bar{D}_{\bullet,M} t_M)$ solves the KKT system (5.3).

In the latter case we obtain $\mathcal{E}(\lambda) = \{M\}$, i.e. $\hat{u} = M$.

Proof of Proposition 5.1. For $\beta = 0$ the conditions of (5.3) simplify to $\lambda \geq t_M$ and the first statement follows immediately. We turn to the second statement where $t_m < \lambda < t_M$ and in which case

$$\bar{D}_{\bullet,M}^T \bar{D} \beta = \bar{D}_{\bullet,M}^T \bar{D}_{\bullet,M} \beta_{M,\bullet}.$$

Therefore, the first equality in (5.3) translates to

$$\begin{aligned} \hat{c}_{M,\bullet} &= \left(\frac{\lambda}{\|\beta_{M,\bullet}\|} + \bar{D}_{\bullet,M}^T \bar{D}_{\bullet,M} \right) \beta_{M,\bullet} \\ &= \left(\frac{\lambda}{(t_M - \lambda)/(\bar{D}_{\bullet,M}^T \bar{D}_{\bullet,M})} + \bar{D}_{\bullet,M}^T \bar{D}_{\bullet,M} \right) \beta_{M,\bullet} \end{aligned} \tag{5.5}$$

which is fulfilled by the definition of $\beta_{M,\bullet}$. Here, the latter equality in (5.5) holds true since the former equality in (5.5) implies

$$t_M = \|\hat{c}_{M,\bullet}\| = \lambda + \bar{D}_{\bullet,M}^T \bar{D}_{\bullet,M} \|\beta_{M,\bullet}\|.$$

For $\lambda \uparrow t_M$ we have $\beta \rightarrow 0$ and therefore $\|G_i\| \rightarrow t_i < t_M$ for $i \neq M$ which yields the assertion. \square

We are now in the position to provide the short proof for Proposition 2.4.

Proof of Proposition 2.4. Straightforward calculations yield

$$(\bar{D}^T \bar{Y})_{i,k} = -w(i, n) \sum_{j=1}^i (Y_{j,k} - \bar{Y}_{n,k})$$

for all $i = 1, \dots, n-1$ and $k = 1, \dots, d$. Hence, we obtain $t_i^2 = \|\hat{c}_{i,\bullet}\|^2 = t(i)$. Since $t(i)$ has a unique maximum, we know that an appropriate λ with $t_m < \lambda < t_M$ may be selected which, together with Proposition 5.1, completes the proof. \square

We continue with the proofs for Subsection 2.2.3.

Proof of Example 2.8. According to (2.30) and due to independence we have

$$\begin{aligned} \text{Var}(S_{i,k}(\varepsilon)) &= \text{Var} \left(\sum_{j=1}^n a_{i,j} (\eta_{j,k} + \phi \eta_{j-1,k}) \right) \\ &\quad + \theta^2 \text{Var} \left(\sum_{j=1}^n a_{i,j} (\eta_{j,k-1} + \phi \eta_{j-1,k-1}) \right). \end{aligned}$$

Straightforward calculations yield

$$\begin{aligned}
& \frac{n}{\tilde{\sigma}^2} \text{Var} \left(\sum_{j=1}^n a_{i,j} (\eta_{j,k} + \phi \eta_{j-1,k}) \right) \\
&= (1 - i/n)^2 [i(1 + \phi^2) + 2(i - 1)\phi] \\
&\quad + (i/n)^2 [(n - i)(1 + \phi^2) + 2((n - i) - 1)\phi] \\
&\quad - 2(1 - i/n)(i/n)\phi \\
&= (1 + \phi^2 + 2\phi) [(1 - i/n)^2 i + (i/n)^2 (n - i)] \\
&\quad - 2\phi [(1 - i/n)^2 + (i/n)^2 + (1 - i/n)(i/n)] \\
&= \alpha(\phi)(i(1 - i/n)) - 2\phi,
\end{aligned}$$

where $\alpha(\phi)$ is set in (2.12) and this implies

$$V^2(i)\xi = \alpha(\phi)((i/n)(1 - i/n)) - 2\phi/n$$

with $\xi = \sigma^2/(\tilde{\sigma}^2(1 + \theta^2))$. □

Proof of Theorem 2.9. Using the notation of (2.7) we consider

$$\begin{aligned}
S_{i,k}(Y) &= n^{-1/2} \sum_{j=1}^i (Y_{j,k} - \bar{Y}_{n,k}) \\
&= n^{-1/2} \sum_{j=1}^i (\varepsilon_{j,k} - \bar{\varepsilon}_{n,k}) + \gamma_k n^{-1/2} \sum_{j=1}^i (\zeta_j - \bar{\zeta}_n) - n^{1/2} H(i, u) \Delta_k \\
&= S_{i,k}(\varepsilon) + \gamma_k S_i(\zeta) + C_k,
\end{aligned}$$

with non-random $C_k = -n^{1/2} H(i, u) \Delta_k$ and where $S_i(\zeta) = n^{-1/2} \sum_{j=1}^i (\zeta_j - \bar{\zeta}_n)$. It holds that

$$S_{i,k}(Y)^2 = |S_{i,k}(\varepsilon) + \gamma_k S_i(\zeta)|^2 + 2C_k(S_{i,k}(\varepsilon) + \gamma_k S_i(\zeta)) + C_k^2$$

and together with the independence of the centered $S_{i,k}(\varepsilon)$ and $S_i(\zeta)$ we get

$$E|S_{i,k}(Y)|^2 = E|S_{i,k}(\varepsilon)|^2 + \gamma_k^2 E|S_i(\zeta)|^2 + C_k^2.$$

Due to part 2 of Assumption 2.5 and due to Assumption 2.6 it holds that $\sum_{k=1}^d E|S_{i,k}(\varepsilon)|^2/d = V^2(i)\sigma^2$ and that $\sum_{k=1}^d \gamma_k^2/d = o(1)$, as $d \rightarrow \infty$. Therefore, we get that, as $d \rightarrow \infty$,

$$E\left(\frac{1}{d} \sum_{k=1}^d |S_{i,k}(Y)|^2\right) = V^2(i)\sigma^2 + nH^2(i, u)\tilde{\Delta}_d^2 + o(1) \quad (5.6)$$

with $\tilde{\Delta}_d^2 = \sum_{k=1}^d \Delta_k^2/d$ and for each $i = 1, \dots, n-1$. Now, assume that we already know that

$$\text{Var}\left(\frac{1}{d} \sum_{k=1}^d |S_{i,k}(Y)|^2\right) = o(1) \quad (5.7)$$

for any i . Then, via Chebyshev's inequality, we obtain from (5.6) and (5.7) that, as $d \rightarrow \infty$,

$$\frac{1}{n^2 \tilde{\Delta}_d^2} \frac{t(i)}{d} = \frac{w^2(i, n)}{n \tilde{\Delta}_d^2} \frac{1}{d} \sum_{k=1}^d |S_{i,k}(Y)|^2 \xrightarrow{P} C(i; u, n, r)$$

for each $i = 1, \dots, n-1$. Due to the continuous mapping theorem we obtain

$$\frac{1}{n^2 \tilde{\Delta}_d^2} \frac{\max_{i \notin S} t(i)}{d} \xrightarrow{P} \max_{i \notin S} C(i; u, n, r),$$

where S is set according to (2.18) and which then completes the proof. (We neglect the trivial case of $S^c = \emptyset$.)

It remains to show that (5.7) holds true for any i indeed. It is sufficient to show that all terms

$$\text{Var} \left(\sum_{k=1}^d S_{i,k}^2(\varepsilon) \right), \quad (5.8)$$

$$\text{Var} \left(S_i(\zeta) \sum_{k=1}^d \gamma_k S_{i,k}(\varepsilon) \right) = E(S_i^2(\zeta)) \sum_{k,r=1}^d \gamma_k \gamma_r E(S_{i,k}(\varepsilon) S_{i,r}(\varepsilon)), \quad (5.9)$$

$$\text{Var} \left(S_i^2(\zeta) \sum_{k=1}^d \gamma_k^2 \right) = \left(\sum_{k=1}^d \gamma_k^2 \right)^2 \text{Var} \left(S_i^2(\zeta) \right), \quad (5.10)$$

$$\text{Var} \left(\sum_{k=1}^d C_k S_{i,k}(\varepsilon) \right) = \sum_{k,r=1}^d C_k C_r E(S_{i,k}(\varepsilon) S_{i,r}(\varepsilon)), \quad (5.11)$$

$$\text{Var} \left(S_i(\zeta) \sum_{k=1}^d \gamma_k C_k \right) = \left(\sum_{k=1}^d \gamma_k C_k \right)^2 \text{Var} \left(S_i(\zeta) \right) \quad (5.12)$$

are of order $o(d^2)$ because the mixed covariance terms can be neglected due to the Cauchy-Schwarz inequality. The fourth moments of any ζ_i are finite, hence $E(S_i^2(\zeta))$ and $\text{Var}(S_i^2(\zeta))$ are finite too. The terms (5.8), (5.9) and (5.11) are of order $o(d^2)$ which follows from Assumptions (2.16) and (2.17) if we take

$$\text{Var} \left(\sum_{k=1}^d S_{i,k}^2(\varepsilon) \right) = \sum_{j,l,q,m=1}^n a_{i,j} a_{i,q} a_{i,l} a_{i,m} \sum_{k,r=1}^d \text{Cov}(\varepsilon_{j,k} \varepsilon_{q,k}, \varepsilon_{l,r} \varepsilon_{m,r})$$

and

$$E \left(S_{i,k}(\varepsilon) S_{i,r}(\varepsilon) \right) = \sum_{j,l=1}^n a_{i,j} a_{i,l} \text{Cov}(\varepsilon_{j,k}, \varepsilon_{l,r}),$$

with $a_{i,j}$ defined in (2.31), into account. The term in (5.10) is of order $o(d^2)$ in view of (2.8) and similarly (5.12) is of order $o(d^2)$ via the Cauchy-Schwarz inequality and again due to (2.8). \square

Proof of Theorem 2.11. The critical function $C(\cdot; u, n, r)$ is positive. It has generally a maximum at $i = u$ for all ratios $r > 0$ if and only if

$$1 \leq \frac{C(u; u, n, r)}{C(i; u, n, r)} \quad (5.13)$$

holds true for all $r > 0$, $i = 1, \dots, n-1$. This can only be fulfilled if

$$1 \leq \lim_{r \rightarrow \infty} \frac{C(u; u, n, r)}{C(i; u, n, r)} = \left[\frac{w(u, n)V(u)}{w(i, n)V(i)} \right]^2 \quad (5.14)$$

holds true for every $i = 1, \dots, n-1$ and the positive weights $w(i, n) = \alpha/V(i)$ from (2.21) fulfill these constraints.

If (2.21) does not hold, then either $\alpha \leq 0$ and thus $w(i, n)$ is negative or $w(i, n)V(i)$ is not a constant function in i and thus

$$1 < \left[\frac{w(q, n)V(q)}{w(p, n)V(p)} \right]^2$$

now holds true for some $q \neq p$. The former contradicts the positivity constraint and the latter contradicts condition (5.14) for $u = p$, $i = q$ and therefore also (5.13) for some $r > 0$. \square

Proof of Theorem 2.12. It holds $w^2(i, n)V^2(i) \equiv 1$. Hence, for any $r \geq 0$, a unique maximum of $C(i; u, n, r)$ is equivalent to

$$\begin{aligned} 0 &< C(u; u, n, r) - C(i; u, n, r) \\ &= \begin{cases} w^2(u, n)(u/n)^2(1 - u/n)^2 - w^2(i, n)(i/n)^2(1 - u/n)^2, & i < u, \\ w^2(u, n)(u/n)^2(1 - u/n)^2 - w^2(i, n)(u/n)^2(1 - i/n)^2, & i > u \end{cases} \end{aligned}$$

for all $i \neq u$. Due to the symmetry of $V(i)$ this is equivalent to

$$\frac{V(n-i)}{V(n-u)} = \frac{V(i)}{V(u)} = \frac{w(u, n)}{w(i, n)} > \begin{cases} i/u, & i < u, \\ (n-i)/(n-u), & i > u \end{cases}$$

and the assertion follows. \square

Proof of Lemma 2.13. Set $f_u(i) = V(i)/V(u)$, $g_u(i) = i/u$ and observe that

$$f_u(u) = V(u)/V(u) = 1 = u/u = g_u(u).$$

Now, assume that $f_u(1) = V(1)/V(u) > 1/u = g_u(1)$. Since $f_u(i)$ is strictly concave in i and $g_u(i)$ is linear in i the assertion follows immediately. \square

Proof of Remark 2.14. We have to distinguish the two cases $\alpha(\phi) > 0$ and $\alpha(\phi) \leq 0$. In the first case $V(i)$ is strictly concave and we may use Lemma 2.13. Simple calculations show

$$\begin{aligned} c(V^2(1) - V^2(u)/u^2) &= (un)^{-2}(u-1)((u+u\phi^2-2\phi)n+2\phi u) \\ &= (un)^{-2}(u-1)(\phi^2 nu - 2\phi(n-u) + nu) \end{aligned}$$

for some $c > 0$ and it is easy to check that $V^2(1) - V^2(u)/u^2 > 0$ holds true (for $\phi \geq 0$ we check via the first equality and for $\phi < 0$ via the second equality). The latter case, $\alpha(\phi) \leq 0$, occurs only for negative ϕ with

$$-1 - n^{-1} - (2n^{-1} + n^{-2})^{1/2} \leq \phi \leq -1 - n^{-1} + (2n^{-1} + n^{-2})^{1/2}. \quad (5.15)$$

In this case it is sufficient to check that

$$h(t) := \left(\alpha(\phi) \left((t/n)(1 - t/n) \right) - 2\phi/n \right) / t^2$$

is strictly decreasing on $t \in [1, n-1]$. Hence, $h(i)/\xi = V^2(i)/i^2$ with $\xi = \sigma^2/(\tilde{\sigma}^2(1+\theta^2))$ is strictly decreasing too. It holds that

$$nt^3 \partial_t h(t) = -\alpha(\phi)t + 4\phi$$

and a sufficient condition for $h(t)$ to be strictly decreasing is that $-\alpha(\phi)t + 4\phi < 0$ holds true. This condition is fulfilled for any $t \in [1, n-1]$ whenever it is fulfilled for $t = n$. The latter is equivalent to $n(\phi^2 + 2\phi + 1) > 2\phi$ which always holds true since $\phi < 0$. Finally, $V(i) > 0$ follows immediately from (2.30). \square

Proof of Theorem 2.16. We assume that $u > n/2$, i.e. $u^* = u$ and set $w = w^{\text{weighted}}$. (The case $u < n/2$ follows by symmetry.) We consider the case $i \leq u$ first and define

$$C(x, r) = F(x)r + G(x)$$

with

$$\begin{aligned} F(x) &= w^2(x, n)V^2(x), \\ V^2(x) &= (x/n)(1 - x/n), \\ G(x) &= w^2(x, n)((x/n)(1 - u/n))^2 \end{aligned}$$

on $[0, n]$, i.e. $C(x, r) = C(x; u, n, r)$ for $x \in \{1, \dots, u\}$. For convenience we suppress the dependence on u and n . It is easy to check that $G(x)$ is strictly increasing with $G(0) = 0$, that $F(x)$ is strictly concave on $[0, n]$ and symmetrical with respect to $x = n/2$ and that $F(0) = 0 = F(n)$ holds true. Further, $C(x, r)$, as a function of x , has a unique maximum at u if and only if $C(x, r) < C(u, r)$ for any $x \neq u$. Now, $C(x, r) < (>)C(y, r)$, $x < y$ and $n/2 < y$ is equivalent to

$$r \begin{cases} < (>)R(y, x), & n - y < x < y, \\ > (<)R(y, x), & 0 < x < n - y \end{cases}$$

and simply $G(x) < (>)G(y)$ if $x = n - y$, where

$$R(y, x) = -\frac{G(y) - G(x)}{F(y) - F(x)} \begin{cases} > 0, & n - y < x < y, \\ < 0, & 0 < x < n - y. \end{cases} \quad (5.16)$$

The latter holds true because $G(x)$ is strictly increasing on $[0, y]$, i.e. $G(y) - G(x) > 0$ is strictly decreasing in x , and because of the properties of F described above. In the following we assume that $y > n/2$ and $x < y$. Since $G(x) < G(y)$ we know

that $C(n-y, r) < C(y, r)$ for any r and since F is symmetric we also conclude that $R(y, x) > R(y, n-x)$ for $n-y < x < n/2$. Altogether, this implies that $C(i, r) < C(u, r)$ for all (discrete) $0 < i < u$ and $r \geq 0$ if and only if

$$0 \leq r < \mathcal{R} = \min_{n/2 \leq i < u} R(u, i).$$

The function $C(x; u, n, r)$ is obviously strictly decreasing for $n/2 < u < x < n$ and the claim follows. \square

Proof of Theorem 2.17. We restrict our considerations to $u > n/2$. The case $u < n/2$ follows by symmetry. In [Bleakley and Vert \(2011a, Theorem 2\)](#), i.e. for $\gamma = 0$, it is used that $C(i; u, n, r)$ has a global maximum at u only if $C(u; u, n, r) > C(u-1; u, n, r)$. This does not hold true in the case of $\gamma \in (0, 1/2)$ and a global maximum can differ from u even though $C(u; u, n, r) > C(u-1; u, n, r)$ holds true. However, we will see that this situation cannot occur if $s \in (1/2 + 1/n, B(\gamma)]$.

We use the notation from the proof of Theorem 2.16. As mentioned there, it is sufficient to consider the case $i \leq u$. In this case we know from the proof of Theorem 2.16 that a possible local maximum of $C(x, r)$ for $x \in [0, u]$ can only occur at some $x_{\max} \in [n/2, u]$. Moreover, using basic analysis we know that

$$\lim_{r \rightarrow \infty} \frac{C(x, r)}{C(y, r)} = \frac{F(x)}{F(y)}$$

for any $0 < x, y < n$. Due to strict concavity of F we know that for any $0 < \delta < 1$ it holds that $F(n/2) > F(n/2 \pm \delta)$. That is, for sufficiently large r , a local maximum of $C(x, r)$ occurs within $[n/2 - \delta, n/2 + \delta] \cap [n/2, u] = [n/2, n/2 + \delta]$.

Now, we compute the rescaled first derivative of $C(x, r)$ on $[0, n]$ which will be denoted by

$$P(x, r) := \frac{((x/n)(1-x/n))^{2\gamma+1}}{x} \partial_x C(x, r).$$

$P(x, r)$ can be evaluated to a second order polynomial in x and for $x \in (0, n)$ we know that $\partial_x C(x, r) = 0$ if and only if $P(x, r) = 0$. Furthermore, since $C(0, r) = 0$ and $C(x, r) \rightarrow \infty$ as $x \uparrow n$ for any $r > 0$ we may have, in case of $\partial_x C(x, r) = 0$, either only a saddle point or a maximum and a minimum must occur simultaneously at some $0 < x_{\max} < x_{\min} < n$. We also know from previous considerations that $x_{\max} \geq n/2$.

The discriminant $D(r)$ of $P(x, r)$ is a second order polynomial in r with roots

$$r_{1,2} = \frac{-(2\gamma+2) \pm 4\gamma^{1/2}}{2\gamma-1} (1-u/n)^2.$$

$D(r)$, as a second order polynomial, must be positive for $r > r_2$, where $r_2 \geq r_1$. Recall that, $C(x, r)$ has a local maximum within $[n/2, n/2 + \delta]$ for any $0 < \delta < 1$ and all sufficiently large r . Otherwise, $D(r)$ would be negative for $r > r_2$ and we would have no extrema of $C(x, r)$ in case of large r .

The solution of $P(x, r_2) = 0$ is given by $x^* = nB(\gamma) \in (n/2, n)$ which is unique and therefore must be a saddle point of $C(x, r_2)$. For $r_1 < r < r_2$ a real solution to $P(x, r) = 0$ does not exist and therefore $C(x, r)$ does not have any extrema on $(0, n)$. For $0 \leq r' \leq r_1$ real solutions do exist but the corresponding roots of $P(x, r')$ cannot correspond to a maximum or a minimum of $C(x, r')$ on $(0, n)$ as discussed in the following. Assume that it is not a saddle point, then we would have a maximum and a minimum because they must occur simultaneously at some $n/2 \leq x_{\max} < x_{\min} < n$, i.e. $C(x_{\max}, r') > C(x_{\min}, r')$. This implies $r' \geq R(x_{\max}, x_{\min})$ and therefore $C(x_{\max}, r) > C(x_{\min}, r)$ for any $r > r'$, which contradicts the fact of no extrema for $r_1 < r < r_2$. We did not exclude the possibility of a saddle point at $x_s \in (0, n)$ because for our conclusions it won't cause any problems as long as the function remains strictly increasing on $(0, n)/\{x_s\}$.

Assume that $\partial_x C(x_0, r_0) = 0$ for some $n/2 < x_0 < n$, $r_0 > 0$. For any $\varepsilon > 0$ we have

$$\partial_x C(x, r_0 + \varepsilon) = \partial_x C(x, r_0) + \varepsilon \partial_x F(x)$$

with $\partial_x F(x) < 0$ for $x \in (n/2, n)$ and we know that

$$\partial_x C(x, r_0) \begin{cases} = 0, & x = x_0, \\ > 0, & 0 < x < n, \quad x \neq x_0 \end{cases} \text{ and } x_0 \text{ is a saddle point.}$$

The first equality ensures that $\partial_x C(x_0, r_0 + \varepsilon) < 0$ for any $\varepsilon > 0$ and that $C(x, r_0 + \varepsilon)$ has local extrema $n/2 \leq x_{\max} < x_0 < x_{\min}$. The second inequality ensures for saddle points x_0 that, for any $0 < \delta < 1$ we can find an $\varepsilon > 0$ such that $x_{\max}, x_{\min} \in (x_0 - \delta, x_0 + \delta)$. Moreover, we know that $x_{\max} \downarrow n/2$ and that $x_{\min} \uparrow n$ as $r \uparrow \infty$.

The properties discussed above ensure that given $n/2 + 1 < u \leq nB(\gamma)$ it suffices to compare $C(u; u, n, r)$ and $C(u - 1; u, n, r)$ to decide whether a maximum is at u or not. The remaining assertions follow now by simple analysis. \square

Proof of Proposition 2.18. As before, we restrict our considerations to $u > n/2$ and the case $u < n/2$ follows by symmetry. Using the notation of Theorem 2.16 we consider the quantities $R(u, x)$, with a continuous argument $x \in (n - u, u)$.

The case $\zeta \in (1/2, 3/4]$ is already shown in (2.28) in Theorem 2.17 and we continue with $\zeta \in (3/4, 1)$. If $nB(\gamma) < u < n$, the properties in the proof of the previous theorem ensure that for sufficiently large n , for $\gamma \in (0, 1/2)$, the differentiable function $R(u, x)$ must have a local minimum at $x_{\min} \in [n/2, nB(\gamma)]$ such that $C(u, r) > C(x, r)$ holds true, for any $x < u$ in case of $r < R(u, x_{\min})$ and $C(u, r) < C(x, r)$ holds true for some $x < u$ in case of $r > R(u, x_{\min})$. Some tedious but straightforward calculations for $\gamma = 1/4$ allow us to solve $\partial_x R(u, x) = 0$ explicitly and to identify the minimum at $x_{\min} = n/2 + (n^2 - un)^{1/2}/2$. Since, x_{\min} is not necessarily in \mathbb{N} we consider $R(sn, x_{\min} + \delta)$ for any $\delta \in [-1, 1]$, where the limits in n equal (2.29), and (using the mean value theorem) observe that this convergence is uniform in δ . Therefore, $R(sn, \lfloor x_{\min} \rfloor)$ and $R(sn, \lceil x_{\min} \rceil)$ have the same limits and the assertion follows since $\mathcal{R}(s, 1/4)$ corresponds to the former or the latter for each n . Finally, the smoothness properties follow on applying l'Hôpital's rule. \square

Acknowledgment

The author wishes to thank Prof. J. G. Steinebach for helpful comments and Christoph Heuser for suggestions to the proof of Theorem 2.17. The author is also thankful for the valuable comments and suggestions of the anonymous referees that helped to improve the quality of this paper. This research was partially supported by the Friedrich Ebert Foundation, Germany.

References

- Aue A., Horváth L. (2013) Structural breaks in time series. *J. Time Ser. Anal.*, 34(1):1–16
- Bai J. (2010) Common breaks in means and variances for panel data. *J. Econom.*, 157:78–92
- Bickel P. J., Levina E. (2008) Regularized estimation of large covariance matrices. *Ann. Stat.*, 36(1):199–227
- Bleakley K., Vert J. P. (2010) Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neural Inform. Process. Syst.*, 23:2343–2352
- Bleakley K., Vert J. P. (2011a) The group fused LASSO for multiple change-point detection. *arXiv:1106.4199v1*, 1–25
- Bleakley K., Vert J. P. (2011b) The group fused LASSO for multiple change-point detection. Technical report HAL-00602121, 1–25
- Boyd S., Vandenberghe L. (2004) *Convex Optimization*. Cambridge University Press, New York
- Csörgő M., Horváth L. (1997) *Limit Theorems in Change-Point Analysis*. Wiley, Chichester
- Frick K., Munk A., Sieling H. (2014) Multiscale change point inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(3):495–580
- Hadri K., Larsson R., Rao Y. (2012) Testing for stationarity with break in panels where the time dimension is finite. *Bull. Econ. Res.*, Issue Supplement s1., 64:s123–s148
- Hannah L. A., Dunson D. B. (2013). Multivariate convex regression with adaptive partitioning. *J. Mach. Learn. Res.*, 14:3261–3294
- Harchaoui Z., Lévy-Leduc C. (2008) Catching change-points with LASSO. *Advances in Neural Inform. Process. Syst.*, 20:617–624
- Horváth L., Hušková M. (2012) Change-point detection in panel data. *J. Time Ser. Anal.*, 33:631–648
- Horváth L., Rice G. (2014) Extensions of some classical methods in change point analysis. *TEST*, 23(2):219–255
- Jandhyala V., Fotopoulos S., MacNeill I., Liu P. (2013) Inference for single and multiple change-points in time series. *J. Time Ser. Anal.*, doi:10.1111/jtsa.12035, 1–24
- Joseph L., Wolfson D. B. (1992) Estimation in multi-path change-point problems. *Commun. Stat.-Theory Meth.*, 21:897–913
- Joseph L., Wolfson D. B. (1993) Maximum likelihood estimation in the multi-path change-point problem. *Ann. Inst. Stat. Math.*, 45:511–530
- Kim D. (2014) Common breaks in time trends for large panel data with a factor structure. *Econom. J.*, 17:301–337
- Pešťová B., and Pešta M. (2015). Testing structural changes in panel data with small fixed panel size and bootstrap. *Metrika*, doi:10.1007/s00184-014-0522-8, 1–25
- Torgovitski L. (2015) Panel data segmentation under finite time horizon. Preprint on arXiv:1501.00177v2, 1–31
- Wu W.B., Pourahmadi M. (2009) Banding sample autocovariance matrices of stationary processes. *Statist. Sinica*, 19:1755–1768